Forest Service
U.S. DEPARTMENT OF AGRICULTURE

Geospatial Technology and Applications Center | January 2023

# Bloom Mapper Methods Brief

## Version 3.0

### Background

In 2021, the Geospatial Technology and Applications Center (GTAC) began collaborating with the Office of Sustainability and Climate (OSC) and cooperators in USFS Region 6 (Washington state and Oregon) and Wyoming to employ retrospective and near real-time remote sensing-based automated algal bloom mapping techniques. While these methods were effective at detecting algal blooms in areas where water bodies are generally clear, false positives in water bodies that typically lack clarity compromised outputs' utility.

During FY 2023, cooperators from the Bridger-Teton and Shoshone National Forests collaborated with GTAC to further explore a new mapping method that incorporates available field observations into a machine learning modeling framework to provide modeled values of cyanobacteria cell count and biovolume. The goal of this approach was to mitigate false positives common in previous approaches, and to provide a tool for visualizing the outputs.

This document provides a brief description of the resulting methods and how to access the products.

### Methods

#### Study Area

The study area includes all of Wyoming and the Greater Yellowstone Ecosystem (GYE; Figure 1). While field sample data were only collected from across Wyoming, extending beyond that region allows us to evaluate the robustness of the model in other locations.

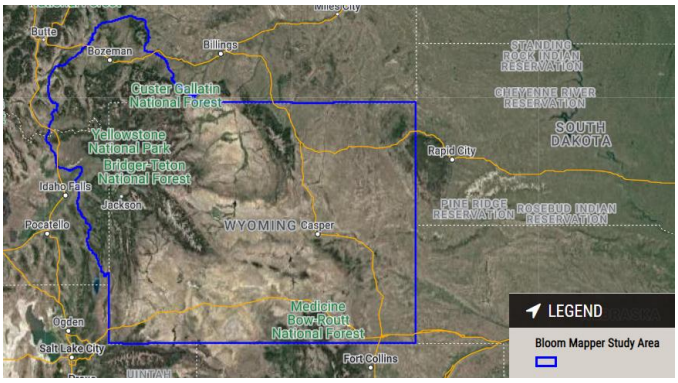


*Figure 1. Bloom Mapper study area that includes the state of Wyoming and the Greater Yellowstone Ecosystem.*

### Computing Environment

We used Google Earth Engine (GEE) (Gorelick 2017) for all earth observation and terrain data access, computation, and output visualization. This is made possible by a USDA Forest Service enterprise agreement with Google. GEE was chosen due to its ability to quickly prototype and test new approaches and then rapidly scale to large-area and multi-temporal applications.

#### Remote Sensing Data Preparation

Copernicus Sentinel- A and B Multispectral Instrument Level-1C (S2) images were accessed through GEE. All S2 images had the S2Cloudless cloud mask applied (where cloud probability > 20; Zupanc 2017), and cloud shadows removed using the Temporal Dark Outlier Mask (TDOM; Housman et al. 2018). All Sentinel-2 data were resampled to 10 m spatial resolution and snapped to the Conterminous United States NLCD USGS Albers Equal Area WGS 84 grid using cubic

convolution. All Sentinel-2 data preparation methods are found within GTAC's GEE data processing and visualization package *geeViz* (https://pypi.org/project/geeViz/; https://github.com/gee-community/geeViz; https://code.fs.usda.gov/forest-service/geeViz;)

### Water Mask Model

Prior iterations of this project demonstrated that up-to-date water masks are needed to avoid false positives on the edge of reservoirs and other water bodies with fluctuating levels. Since reservoirs can significantly fluctuate within a single year, existing water masks were insufficiently up-to-date for near real-time mapping. In 2022, we developed a basic water masking method that identified areas that were dark, wet, and flat (See 2022 Closeout Presentation). This method works well but will commit areas of wet snow/ice, which are then often identified as blooms.

To help mitigate committing wet snow/ice as blooms, we developed a custom supervised water mask model. We then used this model to create a water mask for each Sentinel-2 composite used for both model calibration and application in the forthcoming steps.

We manually collected the training data in a non-systematic manner using basic image interpretation of a late summer image within the Google Earth Engine Playground. We collected a total of 105 snow/ice samples, 170 water samples, and 259 non-snow/ice or non-water samples (notably terrain shadows).

We then used these data with a median Sentinel-2 composite from 2018-2022 August 1-August 16, along with various terrain metrics in a Random Forest classification model (Breiman 2001) using the smileRandomForest function within GEE with 150 trees. The model variable importance can be found in table 1 (See table 2 for a description of the variables).

*Table 1. Water model variable importance. The top variables are dominated by tasseled cap transformation variables and terrain slope.*

| Name | Importance |
|---|---|
| sixth | 22.23 |
| slope | 16.77 |
| tcAngleBG | 15.01 |
| wetness | 13.29 |
| NDVI | 12.54 |
| fifth | 10.88 |
| tcAngleBW | 10.29 |
| greenness | 10.17 |
| hillshade | 9.03 |
| NDMI | 8.96 |
| elevation | 8.64 |
| green | 7.71 |
| brightness | 7.56 |
| NDSI | 7.52 |
| re1 | 7.45 |
| tcDistBW | 7.03 |
| tcDistGW | 6.98 |
| tcDistBG | 6.59 |
| tcAngleGW | 6.37 |
| red | 5.70 |
| blue | 5.57 |
| nir | 5.32 |
| fourth | 5.27 |
| aspect | 5.23 |
| re2 | 4.91 |
| NBR | 4.59 |
| re3 | 4.57 |
| NDCI | 4.34 |
| bloom2 | 3.99 |
| swir2 | 3.98 |
| NDGI | 3.43 |
| tpi_59 | 2.90 |
| nir2 | 2.63 |

| | |
|---|---|
| tpi_29 | 2.36 |
| swir1 | 2.21 |
| waterVapor | 1.82 |
| cirrus | 1.68 |

The out of bag (OOB) model accuracy is 99.4%. However, since the sample was not randomly located, this is not a statistically valid estimate of the accuracy of the water mask.

## Model Calibration Data

Project cooperators provided two types of model calibration data. The first type was field observations from the Wyoming Department of Environmental Quality (WY DEQ) Harmful Cyanobacterial Blooms (HCB) database (https://www.wyohcbs.org/). This database consists of field samples taken over the past several years throughout Wyoming (figure 2). The cyanobacteria cell count (cells/ml) and biovolume ($\mu m^3$) are two attributes we modeled.


*Figure 2. Overview of HCB sample locations throughout Wyoming*

Since HCB data generally have cyanobacteria present, the models also need calibration locations where cyanobacteria are absent. Field experts on the Bridger-Teton National Forest provided a list of water bodies that are known to not experience algal blooms. For each of those water bodies, we first applied the water model (outlined above) to a median Sentinel-2 composite image from August 1-August 16 of each year from 2018-2022. The water extent mask was then buffered inward 3 pixels (30 m) to reduce the likelihood of any water edge contamination of clean samples. We then drew 150 random samples from within the water mask to serve as our cyanobacteria negative model calibration data (figure 3).


*Figure 3. Example clean water body centroid with the resulting water mask in the upper image. The lower image then shows the resulting 150 clean training sample locations*

## Model Training Data Extraction

For each HCB sample, we first computed the median cloud and cloud shadow-free Sentinel-2 composite value from two weeks before and after the sample date. Most HCB samples were located on the edge of water bodies. In order to avoid possible contamination by non-water pixels, we then applied the water mask model to the composite to exclude non-water observations. We then

buffered the water the mask inward 1 pixel (10 m) to further avoid water edge pixels. To mitigate the inclusion of noise and allow for some error in the sample location information, we computed the mean of a 5x5 pixel window of the composite values within the water mask as the value for each sample. Additionally, elevation data were included from the USGS 3DEP 10 m digital elevation model for the sample location (https://www.usgs.gov/media/files/3dep-spatial-metadata-glossary). The complete list of bands that we used in our models can be found in Table 2Table 3. It includes the spectral bands from Sentinel-2, various indices and transformations, and elevation.

*Table 2. Bands used in the cyanobacteria count and biovolume models. Related source information is also provided.*

| Sentinel 2 Band Name | Descriptive Band Name | Algorithm | Band/Index Source |
|---|---|---|---|
| B1 | cb | NA | Sentinel 2 Level-1C |
| B2 | blue | NA | Sentinel 2 Level-1C |
| B3 | green | NA | Sentinel 2 Level-1C |
| B4 | red | NA | Sentinel 2 Level-1C |
| B5 | re1 | NA | Sentinel 2 Level-1C |
| B6 | re2 | NA | Sentinel 2 Level-1C |
| B7 | re3 | NA | Sentinel 2 Level-1C |
| B8 | nir | NA | Sentinel 2 Level-1C |
| B8A | nir2 | NA | Sentinel 2 Level-1C |
| B9 | waterVapor | NA | Sentinel 2 Level-1C |
| B10 | cirrus | NA | Sentinel 2 Level-1C |
| B11 | swir1 | NA | Sentinel 2 Level-1C |
| B12 | swir2 | NA | Sentinel 2 Level-1C |
| NA | NBR | (nir-swir2)/(nir+swir2) | van Wagtendonk et al 2004 |
| NA | NDCI | (re1-red)/(re1+red) | Mishra & Mishra 2012 |
| NA | NDGI | (green-blue)/(green+blue) | Adaptation from Ho et al 2019 |
| NA | bloom2 | green/blue | Ho et al 2019 |
| NA | NDMI | (nir-swir1)/(nir+swir1) | Wilson & Sader 2002 |
| NA | NDSI | (green-swir1)/(green+swir1) | Hall et al 1995 |
| NA | NDVI | (nir-red)/(nir+red) | Rouse et al 1973 |
| NA | brightness | See publication | Crist 1985 |
| NA | greenness | See publication | Crist 1985 |
| NA | wetness | See publication | Crist 1985 |
| NA | fourth | See publication | Crist 1985 |
| NA | fifth | See publication | Crist 1985 |
| NA | sixth | See publication | Crist 1985 |
| NA | tcAngleBG | atan2(brightness,greenness)/pi | Brooks et al 2014 |
| NA | tcAngleBW | atan2(brightness,wetness)/pi | Brooks et al 2014 |
| NA | tcAngleGW | atan2(greenness,wetness)/pi | Brooks et al 2014 |
| NA | tcDistBG | hypotenuse(brightness,greenness) | Brooks et al 2014 |
| NA | tcDistBW | hypotenuse(brightness, wetness) | Brooks et al 2014 |
| NA | tcDistGW | hypotenuse(greenness, wetness) | Brooks et al 2014 |
| NA | elevation | NA | USGS 3DEP 10m |

For each clean sample, the median Sentinel-2 value from 2018-2022 August 1-August 16 was used (the same time period that was used to create the water mask that the clean sample was drawn from). This time period was determined to have a low likelihood of contamination from snow/ice, and therefore more likely to represent a clean water sample.

## Model Training

We used total of 37 HCB samples with a cyanobacteria count > 25000 cells/ml (the WY advisory level threshold) and 3127 clean training points to calibrate two Random Forests regression models (Breiman 2001) – cyanobacteria count and cyanobacteria biovolume. The version of Random Forest within GEE we used is smileRandomForest.

Each model consisted of 150 trees. All other parameters were left at the default. The models' variable importance and OOB error can be found in table 3.

*Table 3 Final random forest models' variable importance and OOB error. For both models, the variables used in earlier phases of this project were the top predictor – bloom2 and NDGI*

| Cell Count | | Biovolume | |
|---|---|---|---|
| Name | Importance | Name | Importance |
| bloom2 | 1.60E+17 | bloom2 | 2.85E+21 |
| NDGI | 1.31E+17 | NDGI | 2.11E+21 |
| wetness | 1.10E+17 | waterVapor | 1.41E+21 |
| waterVapor | 9.01E+16 | tcAngleBG | 1.20E+21 |
| tcAngleBW | 8.22E+16 | tcDistGW | 1.12E+21 |
| sixth | 7.89E+16 | wetness | 1.11E+21 |
| elevation | 6.90E+16 | greenness | 9.12E+20 |
| greenness | 6.46E+16 | sixth | 9.12E+20 |
| NDVI | 6.16E+16 | tcAngleBW | 8.36E+20 |
| tcDistGW | 6.03E+16 | NDCI | 7.38E+20 |
| NBR | 5.98E+16 | re1 | 7.21E+20 |
| NDCI | 5.42E+16 | brightness | 6.93E+20 |
| tcAngleBG | 4.34E+16 | tcDistBW | 6.88E+20 |
| brightness | 4.26E+16 | swir2 | 6.58E+20 |
| NDMI | 4.24E+16 | NBR | 6.54E+20 |
| re1 | 4.24E+16 | NDMI | 6.47E+20 |
| swir1 | 4.04E+16 | swir1 | 5.88E+20 |
| red | 3.86E+16 | red | 5.58E+20 |
| swir2 | 3.26E+16 | re2 | 5.21E+20 |
| fifth | 2.67E+16 | elevation | 5.21E+20 |
| tcAngleGW | 2.63E+16 | NDVI | 4.99E+20 |
| re2 | 2.57E+16 | tcAngleGW | 4.54E+20 |
| tcDistBW | 2.45E+16 | re3 | 4.39E+20 |
| NDSI | 2.38E+16 | nir | 4.30E+20 |
| nir | 2.07E+16 | fourth | 4.12E+20 |
| nir2 | 2.04E+16 | green | 3.19E+20 |
| blue | 1.57E+16 | cirrus | 3.15E+20 |
| cirrus | 1.51E+16 | fifth | 2.60E+20 |
| re3 | 1.45E+16 | nir2 | 2.15E+20 |
| fourth | 1.31E+16 | blue | 2.13E+20 |
| tcDistBG | 1.23E+16 | NDSI | 1.20E+20 |
| green | 1.13E+16 | tcDistBG | 8.50E+19 |
| **OOB Error** | **1,736,840 cells/ml** | | **219,107,057 µm³** |

Notably, the first two variables for both models were the primary variables used in the first two phases of this broader effort. While the OOB error is statistically invalid since the input sample data were not randomly located, it is useful for gaining a sense of how well the model performed at predicting the values at the training sample locations. The OOB error was 1,736,840 cells/ml and 219,107,057µm3 for the cell count and biovolume models respectively. This indicates that the model outputs most likely have a wide error margin. This should be considered when interpreting outputs. For instance, any cell counts output below ~3,000,000 cells/ml is likely error, while cell counts greater than ~6,000,000 cells/ml is likely to have a high cell count. These OOB error results serve as a starting point for lowering the error as more field sample data become available in the future.

## Model Application and Output Delivery

The models were applied to Sentinel-2 4-week median composite images every two weeks for June-October of each year from 2020-2022 for the Wyoming and GYE study area. An output viewer called Bloom Mapper is being tested to provide the ability to view and query outputs.

The development version of Bloom Mapper is located at: https://dev.wrk.fs.usda.gov/forest-atlas/lcms-viewer/bloom-mapper.html while the operational version will be hosted here: https://apps.fs.usda.gov/lcms-viewer/bloom-mapper.html

Currently, we plan to update Bloom Mapper for the 2023 summer season to test the utility of the project outputs.

## Project Resources

### Code

https://code.fs.usda.gov/forest-service/AlgalBlooms

### Past years' presentations:

2021 Closeout Presentation:
https://usfs.app.box.com/file/992261972325

2022 Closeout Presentation:
https://usfs.app.box.com/file/992262027525

### Contacts:

Wyoming USFS Contacts: Gwen Gerber gwen.gerber@usda.gov, Jill McMurray jill.mcmurray@usda.gov

GTAC Technical Contact: Ian Housman* ian.housman@usda.gov

GTAC Leadership Contact: Janet Hsiao janet.hsiao@usda.gov

GTAC Management Contact: Abigail Schaaf abigail.schaaf@usda.gov

*Primary/corresponding author

## References

Breiman, L. (2001). Random Forests. In Machine Learning (Vol. 45, pp. 5-32).

Brooks E.B., Wynne R.H., Thomas V.A., Blinn C.E., Coulston J.W. (2014) On-the-fly massively multitemporal change detection using statistical quality control charts and landsat data. In IEEE Transactions on Geoscience and Remote Sensing (Vol. 52, Issue 6, Article 6573358, pp. 3316-3332). https://doi.org/10.1109/TGRS.2013.2272545.

Crist, E.P. (1985). A TM Tasseled Cap equivalent transformation for reflectance factor data. In Remote Sensing of Environment (Vol. 17, Issue 3, pp. 301-306). https://doi.org/10.1016/0034-4257(85)90102-6.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. In Remote Sensing of Environment (Vol. 202, pp. 18–27). https://doi.org/10.1016/j.rse.2017.06.031

Hall, D.K., Riggs, G.A., Salomonson, V.V. (1995). Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data. In Remote Sensing of Environment. (Vol. 54, pp. 127–140). https://doi. org/10.1016/0034-4257(95)00137-P.

Ho, J.C., Michalak, A.M., Pahlevan, N. (2019). Widespread global increase in intense lake phytoplankton blooms since the 1980s. In Nature (Vol. 574, Article 7778, pp. 667-670.

Housman, I.W., Chastain, R.A., Finco, M.V. (2018). An evaluation of forest health insect and disease survey data and satellite-based remote sensing forest change detection methods: Case studies in the United States. In Remote Sensing (Vol 10, pp. 1184).

Mishra, S., Mishra, D.R. (2014). A novel remote sensing algorithm to quantify phycocyanin in cyanobacterial algal blooms. In Environmental Research Letters (Vol. 9, Article 114003).

Rouse, J.W., Haas, R.H., Schell, J.A.,Deering, D.W. (1973). Monitoring Vegetation Systems in the Great Plains with ERTS (Earth Resources Technology Satellite). In Proceedings of 3rd Earth Resources Technology Satellite Symposium, Greenbelt, 10-14 December. (SP.351, pp. 309-317).

van Wagtendonk, J. W., Root, R. R., & Key, C. H. (2004). Comparison of AVIRIS and Landsat ETM+detection capabilities for burn severity. In Remote Sensing of Environment (Vol. 92, pp. 397−408).

Zupanc, A. (2017) Improving Cloud Detection With Machine Learning. Online: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13. Accessed 20 November 2022.